MACN: A Modality-Aware Cross-Attention Transformer for Non-Invasive Hemoglobin Estimation and Anemia Classification Using Multi-Region Physiological Images

Chaitra S P[1], D R Ramesh Babu[2]
Assistant Professor, Department of Computer Science and Engineering-B.M.S. College of Engineering, Bangalore, Visvesveraya Technological University, Belagavi-590018, India[1]
Professor & Head, Department of Computer Science and Engineering-Dayananda Sagar College of Engineering, Bangalore, Visvesveraya Technological University, Belagavi-590018, India[2]

***Abstract*—Anemia is a major health concern among pregnant women in rural areas. Early detection and timely treatment are necessary to reduce maternal and infant mortality. Many non-invasive techniques have been proposed to solve the problem of unavailability of laboratory-based hemoglobin testing in rural areas. But their accuracy is limited due to dependence on a single modality. This work proposes a multimodal cross-attention network (MCAN) for anemia classification and hemoglobin (Hgb) level estimation using multiple physiological visual cues. Deep convolutional features extracted from four regions of the eye conjunctiva, fingertip, palm, and lip mucosa are fused using a cross-attention multimodal fusion. The fused features are used to estimate Hgb level and predict anemia. Through experimental analysis, the proposed solution is found to increase the estimation accuracy by 5% compared to existing single approaches. The statistical significance of the performance gain is validated through bootstrap confidence intervals and DeLong's test.**

***Keywords*—Anemia; hemoglobin; deep learning; cross-attention fusion; multimodality**

I. Introduction

Anemia is a serious health issue for pregnant women. This nutritional deficiency characterized by lower levels of hemoglobin in the blood. It can increase the chances of maternal and infant mortality when not detected and treated earlier [1]. Thus, early diagnosis becomes very important. A hemoglobin level below 11g/dL is defined as anemia for pregnant women [2]. A clinical method of estimating hemoglobin level from blood samples is a painful, uncomfortable process, as sometimes repeated measurements are needed. The process is time-consuming and infeasible in resource-constrained areas like rural areas. It can also lead to infection spread when facilities are not maintained hygienically [3-4]. Sensing this need for an invasive, comfortable, and safe method for Hgb estimation, various visual cue-based approaches have been proposed. These approaches analyze clinically relevant anatomical sites like eye conjunctiva, fingertip, palm, and lip mucosa, which demonstrate color variations correlated with hemoglobin concentration [5-6]. Various conventional and deep learning techniques have been proposed to estimate Hgb levels from these images. While conventional techniques extracted handcrafted features and trained classification algorithms to estimate Hgb levels, deep learning techniques avoided the need for handcrafted features and applied repeated convolutions and pooling to extract features. Deep learning techniques provided higher accuracy in Hgb estimation in comparison to conventional techniques [7].

The existing deep learning techniques for Hgb estimation can be categorized into single and multi-modality approaches (discussed in the survey). Single-modality approaches use any one of the modalities of eye conjunctiva, fingertip, palm, and lip mucosa for Hgb estimation. Multi-modality approaches use multiple modalities of image or images with clinical metadata to estimate Hgb. Dependence on a single modality image makes the single modality approaches highly sensitive to illumination variations, skin tone differences, etc. Also, individual anatomical regions may not consistently reflect hemoglobin-related color changes. These factors lead to inaccurate hemoglobin estimation [8].

Multimodality approaches integrate complementary information from multiple cues to solve this problem. But current multimodality approaches don't effectively learn cross-modality feature representation from the modalities, leading to reduced accuracy and generalizability of hemoglobin prediction. This work proposes a multimodal cross-attention network to learn robust feature representation, facilitating improved accuracy and generalizability of hemoglobin estimation. The following are the novel contributions of this work:

- A multimodal cross-attention network integrating information from multiple cues of eye conjunctiva, fingertip, palm, and lip mucosa for hemoglobin estimation. Transformer features extracted from each individual modality are fused through a cross-attention fusion technique and the fused representation is used for hemoglobin estimation using a regression head.
- A modality-aware transformer strategy is introduced to extract discriminative color and texture patterns related to hemoglobin concentration using transformer-based encoders. In this way, subtle physiological variations

across different cues are captured while preserving modality-specific information before fusion.

The rest of the study is organized as follows: Section II surveys the existing methods for hemoglobin estimation using deep learning models. Section III details the proposed multimodal cross-attention network integrating multiple cues for hemoglobin estimation. Section IV presents the results of the proposed technique in comparison to recent multimodality and single-modality baselines. Section V presents the conclusion and future scope of work.

II. Related Work

Various single-modality approaches using different variants of deep learning models have been proposed. Jayakody et al. [9] extracted convolutional neural network (CNN) features from fingertip images and used them for estimating the Hgb level. There were no sensitivity and specificity measurements in this work. Magdalena et al. [10] classified eye conjunctiva images into anemic classes using CNN. The image was trained as a whole without consideration for significant regions in the eye conjunctiva image. Das et al [11] combined CNN with vision transformer (ViT) to predict Hgb level from eyelid images. The outliers are more than 8%, and accuracy is very low at the border of anemia Hgb ranges in this approach. Zhang et al. [12] extracted deep learning features from face images and used them for anemia prediction. Authors experimented with ResNet50, MobileNet, InceptionV3, EfficientNetB0, and DenseNet121, and achieved 80% accuracy. But using a face image as a whole without considering the illumination environment can increase the false positives.

Muljono et al. [13] used a hybrid deep learning model combining MobileNetV2 with a support vector machine classifier to predict anemia from eye conjunctive images. Rivero-Palacio et al. [14] classified eye conjunctive image to the anemic class using the YOLOv5 model. The method was tested only for children below 5 years old. Asare et al. [15] used the AlexNet deep learning model to classify anemia. The authors experimented with eye conjunctiva, palm, and fingernail images. Palm images were found to provide higher accuracy, but the performance was tested for children below 5 years old. Magdalena et al. [16] detected anemia from palpebral conjunctiva images using CNN. But the method was tested only for a smaller dataset. Mitani et al. [17] classified retinal fundus images for anemia using the Inception V4 deep learning model. The images were processed as a whole without paying attention to significant regions where color variations are dominant. Chen et al. [18] estimated Hgb level from conjunctiva image using the MobileNet V3 network. The mean absolute error is more than 1.521, and this increases false positives.

Kasiviswanathan et al. [19] proposed a multimodality technique to predict Hgb level. Eye conjunctival image along with age, height, and weight of the subject is used to predict Hgb level. The authors extracted color features from the eye conjunctival image and fused them with meta attributes. A linear regression model is fit between fused features and Hgb level. Image features were not extracted from salient regions. Ramzan et al. [20] combined palm image with corpuscular concentration levels to classify anemia. A deep learning attention model was used to get the fused features. The fused features are classified by softmax classifier to anemia. But the approach needs corpuscular concentration levels which require blood test, making it invasive. Khan et al. [21] extracted fused features from retinal fundus images and meta attributes (age, gender) though a Inception V4 deep learning model and classified the fused features to anemia. But the approach was tested only for people above 50 years old. There were very few works on use of multimodal deep learning.

III. Methodology

Individual modalities of eye conjunctiva, fingernail, palm, and lip mucosa contain useful physiological information related to hemoglobin concentration. Since each modality may be affected by factors such as illumination variations, skin tone differences, or imaging noise, learning cross-modal relationships between these anatomical cues is essential for robust hemoglobin estimation. Though the modalities differ visually, by projecting each modality's features into a shared embedding space cross modal relationship can be learnt. Based on this observation, the architecture of the proposed multimodal cross-attention network (MACN) for Hgb estimation and anemia detection is designed. The architecture is shown in Fig. 1.

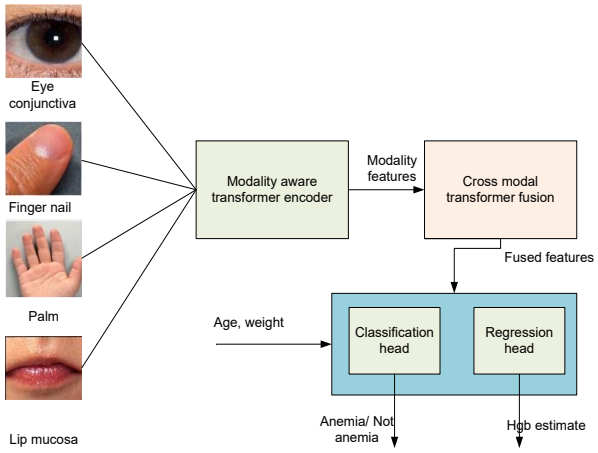

Fig. 1. Architecture of the proposed multimodal cross-attention network.

There are three important functional components in the proposed solution: (1) modality-aware transformer encoder, (2) cross-modal transformer fusion, and (3) classification/regression head. The modality-aware transformer encoder converts the image into a sequence of feature tokens. Cross-modal transformer fusion allows each modality to attend to information from the other modality, resulting in an enhanced, robust feature representation. This feature representation is fused with metadata information (age, weight), and the final fused feature is passed through a classification/regression network. The classification network gives the probability for two classes of anemia/not anemia. The regression head gives the Hgb estimate. Each of the functional components are detailed in the below subsections.

A. Modality-Aware Transformer Encoder

The modality-aware transformer encoder extracts discriminative features from each modality, preserving modality-specific information before fusion. Let the image of modality m be represented as:

$$I_m \in \mathbb{R}^{H\times W\times 3}$$

where, $m \in \{c, f, p, l\}$ corresponds to conjunctiva, fingertip, palm, and lip mucosa, and H, W represent the height and width of the image.

Each input image is split into a sequence of non-overlapping patches of size $P \times P$ with a total number of patches as N. Each patch is flattened and projected into a d-dimensional embedding space:

$$z_0 = \{x_1E, x_2E, \dots x_NE\} \quad (1)$$

where, x_i is the image patch, and E is a learnable embedding matrix.

Positional encoding is added to preserve spatial information:

$$z_0 = z_0 + E_{pos} \quad (2)$$

where, E_{pos} is the positional embedding.

The patch embeddings are processed using a stack of L transformer encoder layers. Each layer has the following components: (1) multi-head self-attention (MHSA), (2) feed-forward network (FFN), (3) normalization layer (NL), and (4) residual connections.

The computation at the l^{th} layer is represented as:

$$z_l' = MHSA\big(NL(z_{l-1})\big) + z_{l-1} \quad (3)$$

$$z_l = FFN\big(LN(z_l')\big) + z_l' \quad (4)$$

The long-range relationships between different patches of the image are learnt using a self-attention operation. For each attention head h, the query (Q), key (K), and values (V) for the attention operation are computed as:

$$Q = zW_Q^h \quad (5)$$

$$K = zW_K^h \quad (6)$$

$$V = zW_V^h \quad (7)$$

The attention weights are computed as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_K}})V \quad (8)$$

The outputs of each attention head are concatenated and projected as:

$$MHSA(z) = concat(head_1, head_2, \dots . head_H)W_O \quad (9)$$

The subtle physiological cues related to hemoglobin concentration are captured by the modality-specific visual representation. For each modality m, its final feature embedding is represented as:

$$F_m = T_m(I_m) \quad (10)$$

where,

$$F_m \in \mathbb{R}^{N\times d} \quad (11)$$

The modality-specific features are passed to the next stage of cross-modal transformer fusion, where inter modal feature relationships are learnt for effective Hgb estimation. The configuration of transformer encoder is given in Table I.

TABLE I. CONFIGURATION OF THE TRANSFORMER ENCODER

Parameter	Value
Input size	224 × 224
Patch size	16 × 16
Embedding dimension	768
Transformer layers	12
Attention heads	12
FFN dimension	3072
Dropout	0.1
Activation	GELU

B. Cross-Modal Transformer Fusion

Cross-modal transformer fusion enables iterative information exchange between modalities through multi-head cross-attention operations, which allows effective feature learning compared to simple feature concatenation. Cross-model transformer allows each modality to attend to information from other modalities. The cross-model interaction between two modalities i and j is represented as:

$$CM(F_i, F_j) = softmax\left(\frac{(F_iW_Q)(F_jW_K)^T}{\sqrt{d}}\right)(F_jW_V) \quad (12)$$

where, W_Q, W_K, W_V are the projection matrices, and d is the embedding dimension. To capture multiple interactions, cross-modal multi-head attention (CMMHA) is used:

$$head_h = Attention(F_iW_Q^h, F_jW_K^h, F_jW_V^h \quad (13)$$

$$CMMHA(F_i, F_j) = Concat(head_1, head_2, .., head_H)W_O \quad (14)$$

where, H is the number of heads and W_O is the output projection matrix.

To capture complementary information across the modalities, bidirectional interactions between the modalities are captured.

$$F_c' = F_c + CMMHA(F_c, F_f) + CMMHA(F_c, F_p) + CMMHA(F_c, F_l) \quad (15)$$

$$F_f' = F_f + CMMHA(F_f, F_c) + CMMHA(F_f, F_p) + CMMHA(F_f, F_l) \quad (16)$$

$$F_p' = F_p + CMMHA(F_p, F_c) + CMMHA(F_p, F_f) + CMMHA(F_p, F_l) \quad (17)$$

$$F_l' = F_l + CMMHA(F_l, F_c) + CMMHA(l, F_f) + CMMHA(F_l, F_p) \quad (18)$$

The modality features are fused to form a cross-fused

representation as in:

$$F_{CF} = Concat(F_c', F_f', F_p', F_l') \quad (19)$$

The cross-fused features capture both modality-specific physiological patterns and cross-modal dependencies. The cross-fused features are passed to the next stage of regression. The configuration of the cross-modal transformer fusion module is given in Table II.

TABLE II. CONFIGURATION OF CROSS-MODAL TRANSFORMER

Parameter	Value
Input modalities	4
Embedding dimension	768
Fusion layer	4
Attention heads	8
FFN dimension	2048
Dropout	0.1
Activation	GELU

C. Classification/Regression Head

Though the cross-fused features extracted from anatomical regions such as eye conjunctiva, fingertip, palm, and lip mucosa provide important information for hemoglobin estimation, clinical attributes, especially age and weight, have a strong influence on hemoglobin levels in pregnant women. Studies have shown that maternal age and anthropometric characteristics such as body weight or body mass index are significantly associated with anemia prevalence and hemoglobin concentration in pregnant women [22-24]. The cross-fused features extracted from visual cues are merged with metadata of age and weight to create final fused features:

$$F_{final} = Concat(F_{CF}, F_{meta}) \quad (20)$$

The final fused representation is passed through a regression network to get Hgb estimation:

$$\overline{Hgb} = R(F_{final}) \quad (21)$$

where, R consists of fully connected layers.

The regression network is trained using the mean square error loss function (L) as:

$$L = \frac{1}{M} \sum_{i=1}^{M} (Hgb_i - \overline{Hgb}_i)^2 \quad (22)$$

where, M is the number of samples, Hgb_i is the actual Hgb value, and $\overline{Hgb}_i$ is the predicted Hgb value.

TABLE III. CONFIGURATION OF THE REGRESSION HEAD

Layer	Operation	Output size
Input	Fused feature vector	F_{final}
FC1	Fully connected + ReLU	256
Dropout	Dropout rate (rate=0.3)	256
FC2	Fully connected + ReLU	128
Dropout	Dropout rate (rate=0.3)	128
FC3	Fully Connected + ReLU	64
Output Layer	Fully Connected (Linear)	1

The configuration of the regression head is given in Table III. The configuration of the classification head is same as that given in Table III, except the last output layer is replaced with the softmax function to provide probabilities for two classes of anemia and not anemia.

IV. RESULTS

The dataset for experimentation was collected from Hadinaru primary health centre (PHC), of rural Mysore district, Karnataka. The data of four images of eye conjunctiva, finger nail, palm, lip mucosa and meta-attributes of age, weight, and Hgb level are collected from 120 participants. Though the dataset consists of 120 participants, each sample contains four physiological image modalities (eye conjunctiva, fingertip, palm, and lip mucosa) together with laboratory-confirmed hemoglobin measurements and clinical metadata. Collecting such multimodal medical data is considerably more challenging than acquiring conventional image datasets because each participant requires synchronized acquisition of multiple anatomical images and corresponding clinical measurements. To ensure reliable evaluation despite the limited dataset size, the proposed model was assessed using 5-fold cross-validation, with results reported as mean ± standard deviation across folds. In addition, bootstrap confidence intervals and DeLong's statistical significance tests were performed to quantify result stability and reliability. The dataset characteristics are given in Table IV.

TABLE IV. DATASET CHARACTERISTICS

Parameter	Value
Dataset acquiring environment	Ambient lighting conditions
Dataset collection duration	6 months
Age range	22-31 years
Weight range	48-85 kg
Hgb range	7 g/dL to 15.5 g/dL
Anemia threshold	11g/dL
Percentage of anemia cases	64 %
Percentage of non-anemia cases	36 %

The dataset is split into training (80%) and test (20%) ensuring balanced distribution of anemic and non-anemic cases in both splits. Testing is conducted with 5-fold cross validation strategy to ensure robustness and generalization. For each fold, a different random split (80:20) is made and the model is trained with training data. The trained model is tested with the test data. The results are reported in terms of mean± standard deviation (SD).

The performance is analyzed in comparison with existing methods specific to each modality, namely fingernail [9], eye conjunctiva [13], palm [15], and multimodal approach [19]. The performance of Hgb estimation is compared in terms of Mean absolute error (MAE), root mean square error (RMSE) and R^2. Anemia classification performance (threshold as 11g/dL) is accuracy and AUC. The performance comparison results with existing approaches are presented in Table V.

The proposed MCAN achieved an accuracy of 95.4% which is atleast 5% higher compared to existing works. The

error in Hgb estimation is at least 41% lower in the proposed MCAN, demonstrating improved precision in Hgb estimation. The lower SD in the proposed MCAN indicates consistent performance across all folds. The single modality approaches ([9],[13],[15]) had higher SD, reflecting sensitivity to imaging conditions and anatomical variability. Though the multimodality approach [19] had lower SD, it lacks robustness in comparison to MCAN. The higher accuracy and lower SD demonstrate the strong generalization capability of the proposed MCAN. The use of modality-aware transformers along with cross-attention fusion has resulted in robust feature representation in the proposed solution. This robust feature representation, along with clinical metadata integration, has enhanced prediction stability and improved estimation accuracy.

TABLE V. PERFORMANCE COMPARISON WITH EXISTING WORKS

Method	Modality used	MAE(g/dL)	RMSE(g/dL)	Accuracy (%)	AUC
Fingernail-based CNN [9]	Fingernail	1.38 ± 0.09	1.79 ± 0.12	83.2 ± 2.3	0.87 ± 0.02
Eye conjunctive CNN [13]	Eye conjunctiva	1.34 ± 0.08	1.71 ± 0.11	84.6 ± 2.0	0.88 ± 0.02
Palm based Deep model [15]	Palm	1.29 ± 0.07	1.65 ± 0.10	86.3 ± 1.8	0.89 ± 0.02
Multi model regression [19]	Eye conjunctiva image + meta data	1.10 ± 0.06	1.42 ± 0.09	90.1 ± 1.6	0.92 ± 0.01
Proposed MCAN	Multi modality image + data	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

The ROC curve of the proposed MCAN is given in Fig. 2.

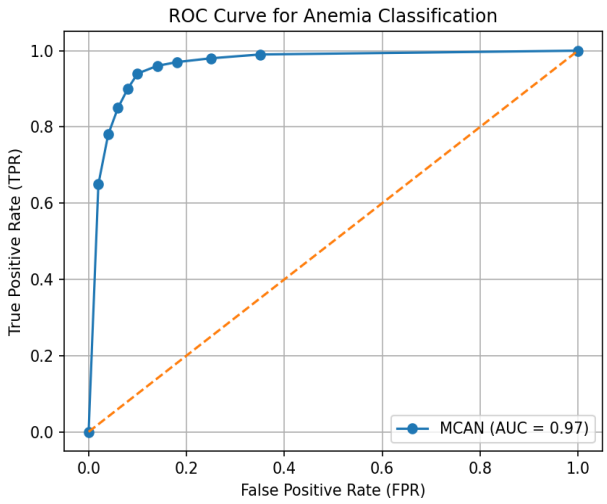

Fig. 2. ROC curve of the proposed MCAN.

The ROC curve shows the excellent discriminative capability between anemia and non-anemia classes in the proposed MCAN. The AUC of 0.97 indicates strong separability. In addition, the position of the curve towards the top left indicates higher sensitivity and specificity across thresholds. The confusion matrix in the proposed MCAN is given in Fig. 3.

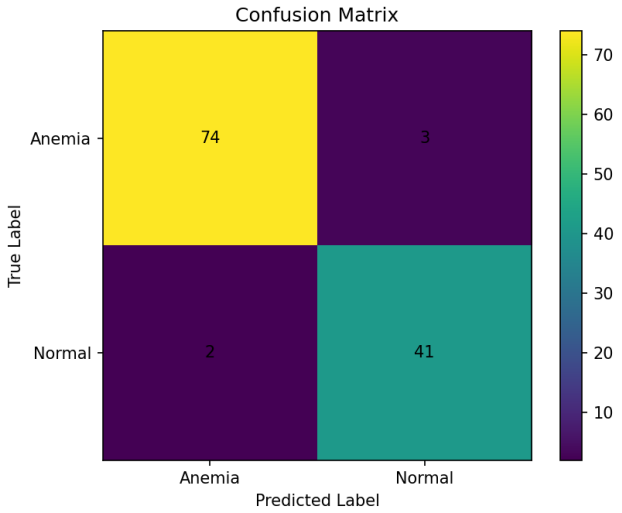

Fig. 3. Confusion matrix

From the confusion matrix, it can be seen that the proposed MCAN has higher sensitivity and specificity, indicating reliable discrimination between anemia and non-anemia classes.

The statistical significance of the performance gain achieved in the proposed MCAN is validated through bootstrap confidence intervals and DeLong's test. The proposed MCAN achieved an accuracy of 95.4% (95% CI: 93.1%–97.2%) and AUC of 0.97 (95% CI: 0.95–0.99). DeLong's test comparing the proposed model with the multimodal solution [19] yielded a p-value of 0.0032, confirming that the improvement is statistically significant.

Cross-attention feature fusion across multiple modalities is an important contribution in this work. Performance comparison is done against different fusion techniques, and the results are presented in Table VI.

TABLE VI. COMPARISON WITH FUSION STRATEGIES

Fusion strategy	MAE (g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
Early fusion	1.18±0.09	1.51± 0.11	89.6 ± 1.9	0.91±0.02
Late Fusion	1.12± 0.08	1.45 ± 0.10	90.7 ± 1.8	0.92 ± 0.02
Feature Concatenation	1.02 ± 0.08	1.31 ± 0.10	91.8 ± 1.7	0.93 ± 0.02
Self-Attention Fusion	0.89 ± 0.07	1.14 ± 0.09	93.4 ± 1.4	0.95 ± 0.01
Proposed MCAN (Cross-Attention Fusion)	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

From the results, it can be seen that the proposed cross-attention fusion achieved best performance in comparison to other fusion techniques. Though self-attention fusion achieved higher performance in comparison to feature concatenation and

early and late fusion, it achieved lower performance in comparison to cross-attention fusion, as it treated all modalities uniformly and did not explicitly model directional interactions between modalities. Cross-attention fusion solved this problem by learning complementary information from other modalities, which produced a more discriminative and robust feature representation.

The performance of the proposed solution is compared against conventional convolutional neural network architectures for feature extraction, and the results are given in Table VII.

TABLE VII. COMPARISON WITH FEATURE EXTRACTION BACKBONES

Backbone	MAE (g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
ResNet50	1.14 ± 0.08	1.47 ± 0.11	89.8 ± 1.9	0.91±0.02
DenseNet121	1.08 ± 0.08	1.39 ± 0.10	90.9 ± 1.8	0.92 ± 0.02
EfficientNet-B0	1.01 ± 0.07	1.30 ± 0.09	92.1 ± 1.6	0.94 ± 0.02
Vision Transformer (ViT-Base)	0.92 ± 0.07	1.18 ± 0.08	93.3 ± 1.5	0.95 ± 0.01
Swin Transformer	0.86 ± 0.06	1.09 ± 0.08	94.1 ± 1.3	0.96 ± 0.01
Proposed MCAN (Modality-Aware Transformer)	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

From the results, it can be seen that traditional CNN architectures – ResNet50, DenseNet121 and EfficientNet-B0 achieved only lower performance compared to transformer-based models. Due to their ability to capture long-range dependencies and subtle color-texture relationships associated with hemoglobin concentration, transformer-based models provided better performance. Compared to standard vision transformers, Swin Transformer provided better results because of its hierarchical feature representation and shifted-window attention mechanism. The proposed MCAN improved the performance even more due to its ability to learn modality-specific physiological characteristics before multimodal fusion. This preserved discriminative information from conjunctiva, fingertip, palm, and lip mucosa images.

Robustness analysis of the proposed MCAN is conducted against different skin tones, and the results are given in Table VIII.

TABLE VIII. ROBUSTNESS ANALYSIS

Skin tone group	MAE (g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
Light	0.76 ± 0.05	0.96 ± 0.07	95.8 ± 1.1	0.97 ± 0.01
Medium	0.79 ± 0.05	0.99 ± 0.07	95.2 ± 1.2	0.97 ± 0.01
Dark	0.83 ± 0.06	1.04 ± 0.08	94.6 ± 1.3	0.96 ± 0.01

Performance evaluation against different skin tones shows that the maximum difference in classification accuracy between skin-tone groups was 1.2%, demonstrating that the proposed multimodal fusion strategy reduces dependence on a single visual cue and improves generalizability across diverse populations.

An ablation study was conducted to analyze the contribution of each functional component of the proposed MCAN. The ablation study results for various variants are given in Table IX.

TABLE IX. ABLATION STUDY RESULTS

Variants	MAE(g/dL)	RMSE (g/dL)	Accuracy (%)	AUC
Without cross-attention (feature concatenation)	1.02 ± 0.08	1.31 ± 0.10	91.8 ± 1.7	0.93 ± 0.02
Without modality-aware transformer (CNN features)	0.95 ± 0.07	1.22 ± 0.09	92.6 ± 1.5	0.94 ± 0.02
Without clinical metadata (age, weight)	0.88 ± 0.06	1.10 ± 0.08	93.9± 1.3	0.95 ± 0.01
Full model (MCAN)	0.78 ± 0.05	0.98 ± 0.07	95.4 ± 1.2	0.97 ± 0.01

Simple feature concatenation without cross-attention fusion reduced the accuracy by 3.6%, demonstrating the robust feature representation effectiveness of cross-attention fusion. Using CNN features instead of modality-aware transformers reduced the accuracy by 3%, indicating the importance of capturing long-range dependencies and subtle color-texture variations with modality-aware transformers. With metadata fusion, the performance degraded only by 1.5%, but still it confirms that physiological attributes complement visual cues. The results of the lowest error, highest accuracy, and lowest SD in the proposed MCAN indicate that each component contributes synergistically to improved hemoglobin estimation and anemia detection.

The novelty of the proposed solution over existing multimodal approaches is given in Table X.

TABLE X. THE NOVELTY OF THE PROPOSED SOLUTION OVER EXISTING MULTIMODAL APPROACHES.

Feature	[19] Kasiviswanathan	[20] Ramzan	Proposed MCAN
Multiple image modalities	✗	✗	✓
Transformer encoder	✗	✗	✓
Modality-aware feature learning	✗	✗	✓
Cross-attention fusion	✗	✗	✓
Metadata integration	✓	✓	✓
Hemoglobin regression	✓	✗	✓
Anemia classification	✓	✓	✓
End-to-end multimodal learning	✗	Partial	✓

The proposed multimodal cross-attention network (MCAN) differs from existing approaches in four important aspects:

1. Differing from prior studies that utilize a single anatomical region or a single image modality, the proposed framework simultaneously exploits four physiologically relevant visual cues, namely eye conjunctiva, fingertip, palm, and lip mucosa. Each modality contains complementary information associated with hemoglobin concentration. Integrating these modalities enables the model to remain robust when one modality is affected by illumination artifacts, pigmentation differences, or image acquisition noise.
2. A modality-aware transformer encoder is introduced for feature extraction. Most of the existing works use convolutional neural networks or generic vision transformers that process all modalities in a similar manner. In contrast, the modality-aware transformer is designed to preserve modality-specific characteristics before fusion, allowing the network to capture subtle color and texture variations associated with hemoglobin concentration while maintaining the unique physiological information present in each anatomical region.
3. The proposed framework uses cross-attention fusion rather than conventional feature concatenation or independent attention mechanisms. Through cross-attention, each modality actively attends to informative regions of other modalities, enabling explicit learning of inter-modal relationships. This interaction produces a richer and more discriminative feature representation that captures both modality-specific and cross-modal dependencies.
4. The proposed work integrates visual information with clinical metadata such as age and weight within a unified prediction framework. While previous multimodal methods generally combine image-derived features with metadata using shallow regression models, the proposed approach learns joint representations through deep multimodal feature interaction. This enables the model to exploit complementary physiological and demographic information for improved hemoglobin estimation.

Thus, the novelty of the proposed MCAN does not arise solely from combining multiple modalities, but from proposing a unified architecture that combines modality-aware transformer encoding, cross-attention-based multimodal fusion, and metadata-guided prediction for non-invasive hemoglobin estimation and anemia detection. To the best of our knowledge, no previous study has jointly incorporated these components within a single end-to-end framework for anemia prediction using conjunctiva, fingertip, palm, and lip mucosa images.

V. Conclusion

This work proposed a novel multimodal cross-attention network (MCAN) for non-invasive Hgb estimation and anemia detection using multiple visual cue modalities. The solution integrated modality-aware transformers with cross-attention fusion to obtain robust feature representation capturing both modality-specific and inter-modal features. The proposed model had atleast 5% anemia detection accuracy and 41% lower Hgb estimation error compared to existing works. Statistical validation using bootstrap confidence intervals and DeLong's test confirmed that the performance gains are significant.

References

[1] Ezzati M, Lopez AD, Rodgers A, Hoorn SV, Murray CJ. Selected major risk factors and global and regional burden of disease Lancet. 2002;360:1347–60

[2] Conrad, M. E. (2011). Anemia. Boston, MA: Butterworths.

[3] N. Mannino et al., "Smartphone app for non-invasive detection of anemia using only patient-sourced photos," *Nature Communications*, vol. 9, 2018.

[4] J. R. White et al., "Noninvasive hemoglobin measurement using digital photographs of the conjunctiva," *Journal of Biomedical Optics*, vol. 21, no. 10, 2016.

[5] Y. M. Chen, S. G. Miaou, and H. Bian, "Examining palpebral conjunctiva for anemia assessment using image processing," *Computer Methods and Programs in Biomedicine*, vol. 137, pp. 125–135, 2016.

[6] T. B. Donmez et al., "Anemia detection through non-invasive analysis of lip mucosa images," *Frontiers in Big Data*, 2023.

[7] A. Al Mahmud et al., "AI-driven anemia diagnosis: A review of advanced deep learning techniques," *arXiv preprint*, 2025.

[8] S. Mythili et al., "Anaemia detection using deep learning by processing conjunctiva, fingernails, palm and tongue images," IRJAEH, 2024.

[9] J. A. D. C. A. Jayakody and E. A. G. A. Edirisinghe, "HemoSmart: A Non-invasive, Machine Learning Based Device and Mobile App for Anemia Detection," 2020 IEEE REGION 10 CONFERENCE (TENCON), Osaka, Japan, 2020, pp. 1401-1406

[10] Magdalena R, Saidah S, Ubaidah IDS, Fuadah YN, Herman N, Ibrahim N. Convolutional neural network for anemia detection based on conjunctiva palpebral images. JTekInform. 2022;3(2):349-354

[11] Das S, Ahamed F, Das A, et al. (July 26, 2024) NiADA (Non-invasive Anemia Detection App), a Smartphone-Based Application With Artificial Intelligence to Measure Blood Hemoglobin in Real-Time: A Clinical Validation. Cureus 16(7): e65442

[12] Zhang A, Lou J, Pan Z, Luo J, Zhang X, Zhang H, Li J, Wang L, Cui X, Ji B, Chen L. Prediction of anemia using facial images and deep learning technology in the emergency department. Front Public Health. 2022

[13] Muljono, S. A. Wulandari, H. A. Azies, M. Naufal, W. A. Prasetyanto and F. A. Zahra, "Breaking Boundaries in Diagnosis: Non-Invasive Anemia Detection Empowered by AI," in IEEE Access, vol. 12, pp. 9292-9307, 2024

[14] Rivero-Palacio M, Alfonso-Morales W, Caicedo-Bravo E. Mobile application for anemia detection through ocular conjunctiva images. In: 2021 IEEE Colombian conference on applications of computational intelligence (ColCACI); May 2021

[15] Asare, Justice & Appiahene, Peter & Donkoh, Emmanuel & Dimauro, Giovanni. (2023). Iron Deficiency Anemia Detection using Machine Learning Models: A Comparative Study of Fingernails, Palm and Conjunctiva of the Eye Images.. 10.22541/au.167570558.82410707/v1.

[16] Magdalena R, Saidah S, Ubaidah IDS, Fuadah YN, Herman N, Ibrahim N. Convolutional neural network for anemia detection based on conjunctiva palpebral images. JTekInform. 2022;3(2):349-354

[17] Mitani A, Huang A, Venugopalan S, et al. Detection of anaemia from retinal fundus images via deep learning. Nat Biomed Eng. 2020;4(1):18-27

[18] Chen YM, Miaou SG, Bian H. Examining palpebral conjunctiva for anemia assessment with image processing methods. Comput Methods Programs Biomed. 2016 Dec;137:125-135

[19] Kasiviswanathan S, Vijayan TB, John S. Ridge regression algorithm based non- invasive anaemia screening using conjunctiva images. J. Ambient Intell. Humaniz. Comput. Oct. 2020

[20] Ramzan M, Sheng J, Saeed MU, Wang B, Duraihem FZ. Revolutionizing anemia detection: integrative machine learning models and advanced attention mechanisms. Vis Comput Ind Biomed Art. 2024 Jul 17;7(1):18

[21] Khan R, Maseedupally V, Thakoor KA, Raman R, Roy M. Noninvasive Anemia Detection and Hemoglobin Estimation from Retinal Images

Using Deep Learning: A Scalable Solution for Resource-Limited Settings. Transl Vis Sci Technol. 2025 Jan 2;14(1):20

[22] L. Balarajan, U. Ramakrishnan, E. Ozaltin, A. Shankar, and S. Subramanian, "Anaemia in low-income and middle-income countries," *The Lancet*, vol. 378, no. 9809, pp. 2123–2135, 2011.

[23] S. S. Stevens et al., "Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in pregnant and non-pregnant women," *The Lancet Global Health*, vol. 1, no. 1, pp. e16–e25, 2013.

[24] N. B. Kozuki et al., "Maternal anthropometry and risk of anemia in pregnancy," *The American Journal of Clinical Nutrition*, vol. 98, no. 2, pp. 429–436, 2013